Struct Chem (2011) 22:1047–1051
DOI 10.1007/s11224-011-9793-0

ORIGINAL RESEARCH

On the relation between basis set convergence and electron correlation: a critical test for modern ab initio quantum chemistry on a "mindless" data set

Roman M. Balabin

Received: 6 January 2011 / Accepted: 12 April 2011 / Published online: 3 May 2011

Abstract The correlation of the only two error sources in the solution of the electronic Schrödinger equation is addressed: the basis set convergence (incompleteness) error (BSIE) and the electron correlation effect. The electron correlation effect and basis set incompleteness error are found to be correlated for all of the molecules in Grimme's "mindless" data set (MB08-165). One can use an extrapolation to the HF or MP2 complete basis set (CBS) limit to see with which type of quantum chemical problem ("simple" and "hard") the researcher is dealing. The origin of the slow convergence of the partial wave expansion can be the Kato cusp condition for electron–electron coalescence. Such an extrapolation is possible for many large molecular systems and would give the researcher an idea about the expected electron correlation level that would lead to the desired theoretical accuracy. In other words, it is possible to use not only the CBS energy value itself but the speed with which it is reached to get extra information about the molecular system under study.

Keywords Basis set superposition error (BSSE) · Basis set incompleteness (BSIE) · Intermolecular interactions · Metal ions

Introduction

The main task of modern ab initio quantum chemistry (QC) is the generation of (approximate) solutions to the Schrödinger equation for different molecular systems: from single atoms to molecular complexes and large biomolecules (peptides, proteins, DNAs, etc.) [1–8]. Today many experimental results can be greatly supported by application of QC theoretical methods [9–13]. Unfortunately, an exact solution to the Schrödinger equation is not possible today for systems of practical importance, with sizes above diatomic molecules, ions, or radicals [14–16]. But modern ab initio quantum chemical methods allow one to systematically increase the accuracy of his/her calculation in order to achieve the desired accuracy, though not without rapidly increasing the computational cost (computational time and computer hardware resources needed) [5, 8, 16].

A simple row of QC methods—HF, MP2, CCSD(T), etc.—is extremely popular today to systematically converge the theoretical prediction of molecular energies or properties in a framework of different computational schemes, starting from Gaussian and Weizmann thermochemistry protocols (G*n*/W*n*, $n = 1$–4) [17, 18] to complete basis set/focal-point analysis (CBS/FPA) schemes [19–22]. The need to overcome basis set incompleteness and systematically approach the complete basis set limit has led to the creation of correlation consistent basis sets: (aug-)cc-pV*x*Z, $x = \mathrm{D} - 6$ [23–25].

There are only two sources of error in the solution of the electronic Schrödinger equation (for molecules composed of light elements, within Born-Oppenhiemer approximation): basis set (BS) convergence error due to incompleteness of any finite basis set and electronic structure method error due to incomplete inclusion of electron correlation (EC) [3–5, 8]. For many years these two error sources were regarded and treated as independent ones [8, 16, 26]. Moreover, the classical scheme of ab initio quantum chemistry, shown in Fig. 1, is valid only if the basis set incompleteness and electron correlation are "orthogonal" to each other. In other words, one can

R. M. Balabin (✉)
Department of Chemistry and Applied Biosciences, ETH Zurich, 8093 Zurich, Switzerland
e-mail: balabin@org.chem.ethz.ch

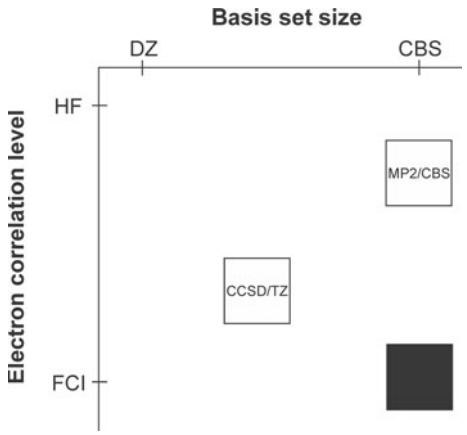

Fig. 1 The approximate solutions of the Schrödinger equation for molecular systems: their ranking in terms of accuracy and convergence to the exact solution (FCI/CBS, *black*). Two perpendicular axes represent basis set (*X*-axis) and electron correlation (*Y*-axis) convergence. The 90° geometry represents the independence of the effects discussed. The common notion in the QC community is that the diagonal "trajectory" is the optimal way to improve the ab initio data

independently increase the electron correlation level *or* basis set and expect to increase the accuracy of the final theoretical answer [16]. However, does it really?

In the coupling between improvement of the basis set and the valence electron correlation method has been studied quantitatively by Martin [26] for the total atomization energies (TAEs) of a number of small molecules (HNO, CO_2, CO, F_2, N_2, N_2O, C_2H_2, CH_4, H_2CO, H_2O, H_2, HCN, HF, NH_3), using basis sets of up to aug′-cc-p-V5Z [7s6p5d4f3g2h/5s4p3d2f1g] quality. Very significant coupling is found to exist [26].

In this communication, we try to test the popular assumption of independence between basis set convergence and electron correlation effect in medium-sized organic molecules. Coupled-cluster and Moller–Plesset theory levels were used together with Dunning correlation consistent basis sets. It was proved that the errors produced by deficiencies in electron correlation treatment and basis set incompleteness are *correlated*.

Calculations

A "mindless" molecular set (MB08-165; 165 items, 8 atoms each) of medium-sized ($M_W = 58 \pm 22$ Da) organic molecules from Prof. Grimme [27] was used throughout the study. Despite being single-reference main-group (H–Cl, without He and Ne) molecules, the generated "artificial molecules" show a large structural diversity with interesting bonding features [27]. Many of the molecules and molecular complexes are, chemically, rather unusual. One expects that this set will provide an unbiased estimation for the whole range of possible organic molecules and even go beyond that [24, 25, 27]. After the initial geometry optimizations at the DFT-PBE/TZVP level and CCSD/cc-pVDZ single-point calculations for the T1- and D1-diagnostics as described in [27], the CCSD(T) complete basis set (CBS) correlation energy was extrapolated using cc-pVTZ and cc-pVQZ single-point values [19, 20, 28]. Note that T1 diagnostics alone might not be fully sufficient: in more thorough analysis one looks also at first few t1- and t2-amplitudes from CCSD since they provide better picture for judging MR cases. The CCSD(T) correction, $\delta\text{CCSD(T)} = E_{\text{CCSD(T)}} - E_{\text{HF}}$, was added to HF/CBS energy, calculated by a three-point scheme (x = D, T, Q). The basis set incompleteness error BSIE was calculated as the difference between the calculated total energy value the finite basis set (e.g., cc-pVDZ) and the CBS energy value: $\text{BSIE} = E_{\text{bs}} - E_{\text{CBS}}$ [3–5]. The δCCSD(T) and BSIE values were normalized by HF/CBS energy or number of electrons (N_{el}) to eliminate the molecular size dependence. The same procedure was applied for MP2, MP3, MP4(SDQ), and MP4/MP4(SDTQ) energies [19, 20]. All coupled-cluster calculations were done using Molpro 2006.1, [27] while Gaussian 09 was used for Moller–Plesset theory application [29].

Of course, a better way to represent the correlation of basis set and electron correlation effects would be to calculate the difference between CCSD(T)/MP4/MP2 and full CI (FCI) results; these would be better coordinates for the representation of a theory level in Fig. 1. Unfortunately, the FCI energy values are available nowadays only for the smallest molecular systems, and the calculation at this level for medium-sized organic molecules is not expected to be possible in the near future [14–16]. Nevertheless, δCCSD(T) can be regarded as a good approximation of electron correlation effects in molecular systems [19–22]. In many cases, the corrections above the CCSD(T) level are negligible [22, 24, 25, 28].

Results and discussion

Figure 2 represents the results of simple statistical analysis of 165 molecules in terms of BSIE vs. δCCSD(T) correlation. The double-log representation allows one to cover a wide range of energy values and shows the data correlation more clearly. From Fig. 2, one can clearly see that a high correlation ($R^2 > 0.84$) between basis set convergence and the electron correlation effect is observed. The degree of this correlation decreases when going from double-ζ ($R^2 = 0.93$) to triple-ζ (0.91) to quadruple-ζ (0.84) basis sets. This can be explained by the saturation effect in the basis set increase when going from an already rather large cc-pVTZ basis set to a cc-pVQZ one [22, 27, 28]. At the

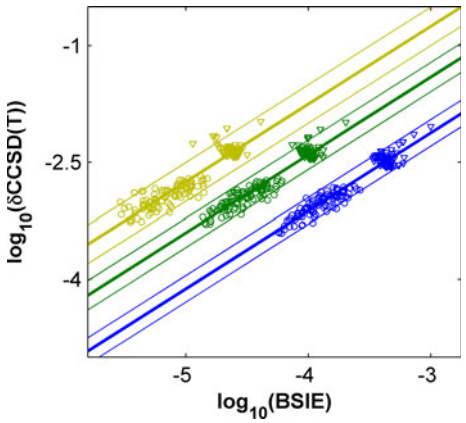

Fig. 2 The correlation between the basis set incompleteness error (BSIE) and correction for electron correlation calculated at the CCSD(T) level. Linear regression models (*thick lines*) are shown together with their 95% confidence intervals (*thin lines*) for cc-pVDZ (*blue*), cc-pVTZ (*green*), and cc-pVQZ (*yellow*) basis sets. The corresponding R^2 values are 0.93, 0.91, and 0.84. *Circles* represent molecules made of second-period atoms only; *triangles* represent molecules containing at least one third-row atom (Color figure online)

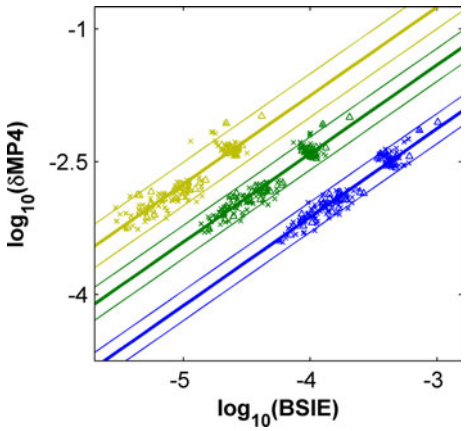

Fig. 3 The correlation between the basis set incompleteness error (BSIE) and correction for electron correlation calculated at the MP4 level. Linear regression models (*thick lines*) are shown together with their 95% confidence intervals (*thin lines*) for cc-pVDZ (*blue*), cc-pVTZ (*green*), and cc-pVQZ (*yellow*) basis sets. The corresponding R^2 values are 0.93, 0.91, and 0.85. Crosses represent closed-shell artificial molecules ($S = 0$, $M = 1$); *triangles* represent open-shell artificial molecules ($S = ½$, $M = 2$) (Color figure online)

Hartree–Fock level, the difference between cc-pVQZ and cc-pVTZ energies is just -11 ± 5 mE_h or -7 ± 3 kcal mol^{-1} ($\pm\sigma$). The same value for the CCSD(T) electron correlation level is -43 ± 19 mE_h or -27 ± 12 kcal mol^{-1}. The linearity of the δCCSD(T)-BSIE dependence is seen especially well in the cc-pVDZ case, where all of the points are inside the model error bars or very close to them [30, 31].

Two groups of points are clearly seen for each basis set in Fig. 2. The first group, with the molecules forming a very good regression pattern, represents second-period molecules (H–F). The second group, with the points crowding in the top-right corner of the plot, represents the molecules with at least one third-period atom (Na–Cl). In the latter case, some points are coming out of the model's 95%-confidence interval, especially in the cc-pVQZ case. This can be explained by the presence of 10 core electrons in the third-row atoms and by the use of the frozen-core approximation for calculating the energy values [32]. Both of these facts make uniform data normalization a complicated task, so the fact that these data points are still inside the model error bars should be regarded as remarkable. It shows the generality of the correlation between basis set and electron correlation effects in organic molecules.

The use of other electron correlation levels, e.g., MP4 as shown in Fig. 3, does not change the results significantly. An almost identical correlation pattern can be well observed in Fig. 3. The same separation into two groups of molecules is observed. Figure 3 also shows that no influence of molecular spin (multiplicity) is found for the data set discussed. So, the correlation between BSIE and δE is the same for open- and closed-shell molecular systems.

The data points in Figs. 2 and 3 were fitted with just one independent parameter (k) of the form:

$$\delta E = k \times \text{BSIE} \quad (1)$$

where k depends on the size of the basis set applied.

Although the exact mathematical representation of this correlation is a separate and complicated question, here we state only that *there is* a dependence between energy and basis set corrections: $\delta E = f(\text{BSIE})$. When this fact is accepted, we can clearly see that outdated and often persisting idea about the accuracy of ab initio quantum chemistry should be corrected. One, of course, can still independently choose the type of electron correlation treatment and basis set size, but he or she should understand that the "molecular" reality is different.

First of all, here we prove that there are "simple" and "hard" problems for ab initio QC: the first class has a quick convergence in both basis set and electron correlation spaces, while the second class has a slow convergence, again, for both quantities [22, 27]. One can even use a simple (with today's hardware) extrapolation to the HF/CBS limit to see with which type of problem he or she is dealing [19, 20]. Such an extrapolation is possible for many large molecular systems and would give the researcher an idea about the expected electron correlation level that would lead to the desired accuracy. In other words, one will now use not only the CBS energy value itself but the speed with which it is reached to get extra information about the molecular system under study. Note that Figs. 2 and 3 also prove that there is no well-defined border between "simple" and "hard" computational problems, and the whole spectrum of problems is expected [26].

Second, the application of rather sophisticated algorithms of electron correlation evaluation, e.g., CCSD(T) method, together with small basis sets (4-31G, 6-31G(d), 6-31G(d,p), D95/D95V, DZV, cc-pVDZ, etc.) cannot be recommended for estimation of "higher-order correlation corrections" [27, 28, 33, 34]. If the CCSD(T) electron correlation level is needed to get a precise energy value, a large basis set is also needed and vice versa. To date, many reported energy differences or other molecular properties are extrapolated from HF/CBS or MP2/CBS values using δCCSD(T) correction with a double-ζ basis set [33, 34]. This practice can lead to an underestimation of the electron correlation effect, although the term "CCSD(T)/CBS" is used for such calculations. Once again, if one really wants to have a better approximation (a solution that is closer to the exact one), he or she needs to increase both the electronic structure theory level and the basis set. Of course, extrapolation to CCSD(T)/CBS theory level is possible, but a large (at least cc-pVTZ) basis set is needed for the CCSD(T) calculation [27, 28]. These data also clarify the low accuracy of CCSD(T)/cc-pVDZ reaction energies for MB08-165 molecules as well as similar facts in other benchmark studies [27].

The exact form of the dependence between basis set convergence and the electron correlation effect is difficult to determine if one wants to be sufficiently general and include different types of molecules and heavy atoms [35]. Alternative to the data in Fig. 2, one can normalize the energy values to the number of electrons in the molecular system calculated (Fig. 4). Unfortunately, although such a normalization scheme allows for the formation of the two groups of molecules, as discussed above, closer on the δE-BSIE plot, some separation between them is still observed (compare squares and triangles in Fig. 4). As expected, another normalization scheme leads to a different form of the dependence between energy correction and basis set convergence:

$$\delta E/N_{\mathrm{el}} = K \times (\mathrm{BSIE}/N_{\mathrm{el}})^{2/3} \quad (2)$$

where K is the only independent parameter. The power dependence seems to be an artifact of not fully removing the size effect.

The use of the total number of electrons in the system as a normalization factor also degrades the quality of the regression, leading to variance inflation factors (VIF) of 2.3 ± 0.4 in contrast to 11 ± 3 for Eq. 1 [36–38]. This is understandable because the HF energy represents the size of a molecular system from a quantum chemical point of view better than just the number of electrons.

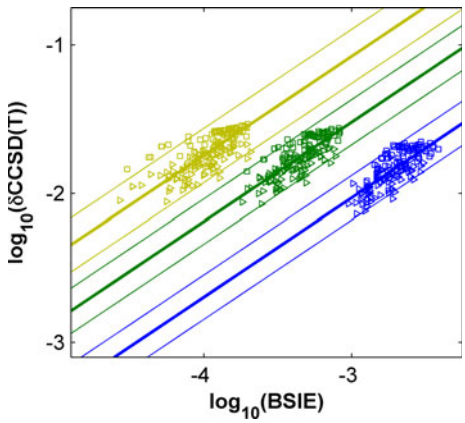

Fig. 4 The correlation between the basis set incompleteness error (BSIE) and correction for electron correlation calculated at CCSD(T) level, normalized by the total number of electrons in molecular system (N_{el}). Linear regression models (*thick lines*) are shown together with their 95% confidence intervals (*thin lines*) for cc-pVDZ (*blue*), cc-pVTZ (*green*), and cc-pVQZ (*yellow*) basis sets. The corresponding R^2 values are 0.61, 0.61, and 0.45. *Triangles* represent molecules made of second period atoms only; *squares* represent molecules containing at least one third-row atom (Color figure online)

Conclusions

From the results presented above, one can conclude that (i) the electron correlation effect and basis set incompleteness error (BSIE) are correlated; [39–42] (ii) the form of their dependence differs with the type of energy normalization; (iii) one can use an extrapolation to the HF/CBS limit to see the type of QC problem ("simple" and "hard") he or she is dealing with [27]. The standard representation of the hierarchy of quantum chemistry algorisms as points in Fig. 1 with two independent coordinates—electron correlation level and basis set size—should be altered [16, 26]. The origin of the slow convergence of the partial wave expansion can be the Kato cusp condition for electron–electron coalescence. Further research can clarify the type of δE-BSIE dependence for charged species, biopolymers (as well as other macromolecules and molecular clusters of scientific and industrial importance), and molecules containing heavy atoms (e.g., transition metals) [43–50]. The type of correlation between basis set convergence and the electron correlation effect for DFT methods is also of great interest [51–54].

Acknowledgments BRM wishes to thank I. Samoilenko for his computational assistance and E. Lomakina for her help in manuscript preparation. S. Grimme and M. Korth are acknowledged for granting of the molecular set data and CCSD(T) energies.

References:

1. Jensen F (1996) Chem Phys Lett 261:633
2. Boys SF, Bernardi F (1970) Mol. Phys. 19:553
3. Balabin RM (2010) J Chem Phys 132:231101
4. Balabin RM (2010) J Chem Phys 132:211103
5. Helgaker T, Klopper W, Koch H, Noga J (1997) J Chem Phys 106:9639

6. Balabin RM (2010) J Phys Chem A 114:3698
7. Simon S, Duran M, Dannenberg JJ (1996) J Chem Phys 105:11024
8. Dunning TH (2000) J Phys Chem A 104:9062
9. Balabin RM (2010) Phys Chem Chem Phys 12:5980
10. Balabin RM (2010) J Phys Chem Lett 1:20
11. Balabin RM (2009) J Phys Chem A 113:4910
12. Fedorov A, Moret M-E, Chen P (2008) J Am Chem Soc 130:8880
13. Balabin RM (2009) J Phys Chem A 113:1012
14. Li Z, Abramavicius D, Mukamel S (2008) J Am Chem Soc 130:3509
15. Booth GH, Alavi A (2010) J Chem Phys 132:174104
16. Jensen F (1999) Introduction to computational chemistry. John Wiley & Sons, Chichester
17. Curtiss LA, Redfern PC, Raghavachari K (2007) J Chem Phys 126:084108
18. Karton A, Rabinovich E, Martin JML, Ruscic B (2006) J Chem Phys 125:144108
19. Balabin RM (2008) Chem Phys 352:267
20. Balabin RM (2009) Chem Phys Lett 479:195
21. Moran D, Simmonett AC, Leach FE, Allen WD, Schleyer PV, Schaefer HF (2006) J Am Chem Soc 128:9342
22. Csaszar AG, Allen WD, Schaefer HF (1998) J Chem Phys 108: 9751
23. Dunning TH (1989) J Chem Phys 90:1007
24. Bachrach SM (2007) Computational organic chemistry. Wiley-Interscience, New York
25. Balabin RM (2009) J Chem Phys 131:154307
26. Martin J (1007) Theor Chem Acc 97:227
27. Korth M, Grimme S (2009) J Chem Theory Comput 5:993
28. Balabin RM (2008) J Chem Phys 129:164101
29. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Mennucci B, Petersson GA, Nakatsuji H, Caricato M, Li X, Hratchian HP, Izmaylov AF, Bloino J, Zheng G, Sonnenberg JL, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Vreven T, Montgomery JA, Jr, Peralta JE, Ogliaro F, Bearpark M, Heyd JJ, Brothers E, Kudin KN, Staroverov VN, Kobayashi R, Normand J, Raghavachari K, Rendell A, Burant JC, Iyengar SS, Tomasi J, Cossi M, Rega N, Millam JM, Klene M, Knox JE, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Martin RL, Morokuma K, Zakrzewski VG, Voth GA, Salvador P, Dannenberg JJ, Dapprich S, Daniels AD, Farkas Ö, Foresman JB, Ortiz JV, Cioslowski J, Fox DJ (2009) Gaussian 09, Revision A.1. Gaussian, Inc., Wallingford, CT
30. Balabin RM, Safieva RZ (2007) J Near Infrared Spec 15:343
31. Balabin RM, Syunyaev RZ (2008) J Colloid Interface Sci 318:167
32. Hosteny RP, Dunning TH, Gilman RR, Pipano A, Shavitt I (1975) J Chem Phys 62:4764
33. Jurecka P, Sponer J, Cerny J, Hobza P (2006) Phys Chem Chem Phys 8:1985
34. Pittner J, Hobza P (2004) Chem Phys Lett 390:496
35. Halkier A, Helgaker T, Jørgensen P, Klopper W, Koch H, Olsen J, Wilson AK (1998) Chem Phys Lett 286:243
36. Hair JF, Anderson R, Tatham RL, Black WC (2006) Multivariate data analysis. Prentice Hall, Upper Saddle River, NJ
37. Balabin RM, Safieva RZ, Lomakina EI (2008) Chemometr Intell Lab Syst 93:58
38. Balabin RM, Safieva RZ, Lomakina EI (2007) Chemometr Intell Lab Syst 88:183
39. Neogrady P, Medved M, Cernusak I, Urban M (2002) Mol Phys 100:541
40. Bartlett RJ, Musial M (2007) Rev Mod Phys 79:291
41. Dedikova P, Pitonak M, Neogrady P, Cernusak I, Urban M (2008) J Phys Chem A 112:7115
42. Pitonak M, Neogrady P, Rezac J, Jurecka P, Urban M, Hobza P (2008) J Chem Theory Comput 4:1829
43. Chen Y-F, Dannenberg JJ (2006) J Am Chem Soc 128:8100
44. Stone AJ (2008) Science 321:787
45. Balabin RM, Lomakina EI (2011) Analyst 136:1703
46. Balabin RM, Safieva RZ (2011) Anal Chim Acta 689:190
47. Holroyd LF, van Mourik T (2007) Chem Phys Lett 442:42
48. Balabin RM, Safieva RZ (2008) Fuel 87:2745
49. Balabin RM, Safieva RZ (2008) Fuel 87:1096
50. Carter EA (2008) Science 321:800
51. Balabin RM, Lomakina EI (2009) J Chem Phys 131:074104
52. Suzuki S, Green PG, Bumgarner RE, Dasgupta S, Goddard WA, Blake GA (1992) Science 257:942
53. Balabin RM, Safieva RZ, Lomakina EI (2011) Microchem J 98:121
54. Balabin RM, Lomakina EI, Safieva RZ (2011) Fuel 90:2007